Date: feb 5, 2004 Express Mail Label No. EV 214951094 US

Inventor: Timothy G. Nye

Attorney's Docket No.: 3014.1007-001

SYSTEM AND METHOD FOR IDENTIFYING AN INTERNET RESOURCE ADDRESS

RELATED APPLICATION

This application claims the benefit of U.S. Provisional Application No. 60/444,874, filed on February 5, 2003. The entire teachings of the above application is incorporated herein by reference.

BACKGROUND

The Internet has become a major source for valuable information relating to products and services available for sale. For example, in the case of the restaurants, menus are often accessible on-line. Despite the fact that the Internet provides such a wealth of information, a problem exists in that it is typically difficult for a user to locate the website of a business even if the exact name and city location of the business is known and used.

For example, a user sees a TV commercial for a restaurant in the city of Boston called "Bertucci's" and wants to visit the website of "Bertucci's" to obtain more information. The user enters the keywords "Boston Bertucci's" into a web search engine, such as the one at www.yahoo.com. The user may receive, for example, a list of 876 matches, but find that the actual Uniform Resource Locator (URL) for the restaurant is not anywhere in the search results. Sometimes the desired match may be returned but buried so deeply in the search results that the user is unable to find the match even if they have the patience to sift through the entire search result list. Further, if the user interface is a Voice Over IP (VoIP) interface, where the search results are audibly read back to the user, the sifting process may take hours and therefore, for most purposes is impractical. There are directories or portals on the Internet that maintain databases relating to specific content such as for example a database of restaurants, for searching by users. Users may query these databases for a more manageable set of search results. However, the Internet is a fluid and dynamic medium where the available information is consistently being edited and expanded. After data has been collected for these databases, the data soon becomes stale as new data is published. Further, in some cases, these databases are still too large yielding search result lists that are too long.

As will be appreciated a problem exists in that there is no reliable method for users to find the website of a particular business or entity on the Internet. Search engines are hit and miss, and yield an overwhelming amount of false positive hits requiring users to spend significant amounts of review time in order to locate the correct website address. Directories or portals with databases of specific information do not provide much of an improvement. This is assuming that there is a directory or portal having the desired subject matter with the website addresses.

Outside of the Internet, users may call businesses to ask for their website addresses, but this only works when the businesses are open. From a business point of view, this process takes time and money to provide the requested information. Further, calling businesses is not always reliable as callers are frequently passed to automated attendants.

Another source of business information is the Yellow Pages, but website addresses are not usually provided except in some of the advertisements. Also, the problem of staleness is even worse with printed media as compared to information available on the Internet.

It is therefore an object of the present invention to provide a novel system and method for locating the web address of an entity based on an attribute of the entity.

SUMMARY

The present invention relates to a method and system for generating highly targeted searches. Preferably, the invention is used to identify a URL address for an entity. An URL address for the entity may be determined based on information known about the entity, such as an attribute of the entity. Computational techniques and prediction processes may be used during a search routine to eliminate false positives and determine a candidate address of the entity.

In one embodiment, an attribute of an entity, such as a business's telephone number, may be used to determine another attribute of the business, such as the business's URL address. In this example, a telephone number may be submitted to one or more search engine, and in response, a list of URL addresses may be generated. Web content may be collected from the website located through the URL address. Alternatively, indexed content associated with the URL address that has been provided by the search engine may be used. The content may be parsed to locate a URL address or email address. The number of times a unique URL address appears throughout all content parsed is computed. If the computed value is above a threshold value, the URL may be an accurate address. A process is performed to eliminate false positives in addresses identified. The URL address that has the highest value may be considered the correct URL address for the entity. The URL addresses determined to be correct may be used to update a directory in an ongoing manner.

In another embodiment, highly targeted searches may be achieved by cross-referencing and narrowing search results using a collection of information. Specialized filtering and parsers may be used to narrow search results. A collection of information may, for example, be any collection of information about businesses.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, features and advantages of the invention will be apparent from the following more particular description of preferred embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating the principles of the invention.

Figure 1 shows the architecture of a search system according to an embodiment of the present invention.

Figure 2A shows a process for locating a website address of an entity based on an attribute of the entity according to an embodiment of the present invention.

Figure 2B shows a process for creating a database of directory websites that include directories, news sites, or portals.

Figure 3 shows a graph of a sample matrix generated in accordance with the process shown in Figure 2B.

Figures 4A and 4B show a process for locating a website address of an entity based on an attribute of the entity in accordance with the present invention.

Figure 5 shows a graph of hits versus URL of a sample search result.

Figure 6 shows a process for using a database as a filter for a search routine according to an embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

A description of preferred embodiments of the invention follows.

System Architecture

Preferably, the invention is implemented in a software or hardware environment. One such environment is shown in Figure 1. In this example, a system 10 is provided for generating highly targeted searches. Although the search technique may be implemented a search engine, it may be desirable to provide a search routine 5, which searches the web 55 using, among other things, existing search engines 15, such as Google or Yahoo. Websites identified by the search engines 15 may be spidered 20 to collect relevant information.

The system 10 may also use a collection of information 25 to optimize searching. The collection of information 25 may be any database, such as a yellow pages database, Better Business Bureau membership list, AARP membership list, etc. A particular collection of information 25 may be provided or selected by a user at a query interface 30. Alternatively, the system 10 may determine an appropriate collection of information 25 to use during the search based on the content of the user's search query.

In performing a search analysis 5, the system 10 may use a variety of processing techniques. These processing techniques may include distillation logic 35, prediction logic 40, domain name logic 45 and parsers 50. The distillation algorithm 35, for example, may be used to eliminate false positives from search results. The prediction algorithm 40 may be used to predict which URLs identified during searches are likely to be accurate. The domain analyzer 45 may be used to analyze domain names in URL addresses that have been identified during searches. At least one parser 50 may by the system 10 to target the user's search query to specific context.

Search Routine

Figure 2A shows a process for locating the website address of an entity based on an attribute of the entity in accordance with one embodiment of the present invention. The process may be implemented in software or hardware. The process involves obtaining a telephone number of the entity of interest (step 105); submitting the telephone number to one or more web based search engines (step 110); receiving a list of URLs from the search engines (step 115); retrieving from each website linked to the URLs, web content (step 120); parsing the retrieved web content to identify email and website addresses therein (step 125); counting each unique website address identified either by an email address or a website address (step 130); and determining the website address of the entity based on the count values according to a prediction algorithm (step 135). It will be understood by those skilled in the art that it is assumed that a website address is incorporated into each email address.

In one preferred embodiment, a telephone number is used as a keyword for submitting to the one or more search engines. Alternately, keywords based on other information such as address, business name, or combinations, including telephone numbers, thereof may be submitted to the search engines.

A prediction algorithm is used to determine which website address has the most hits as a match for the website address of the entity of interest. In the case where a plurality of unique website addresses have the same most number of hits, the prediction algorithm deems all such website addresses to be matches for the website of the entity of interest.

The following is a search example in accordance with distillation algorithm. A user enters a telephone number of a business into one or more webbased search engines to locate the website address for the business. Where a telephone number is not available, a business name may be entered for lookup on a Yellow Page database to obtain the telephone number of the business. The telephone number is then submitted to the search engines (such as for example Google, MSN, alltheweb) with quotes around the telephone number where the search engine requires such. The use of quotations provides slightly better search results on some of the search engines.

From the search result hits returned by the search engines, the URLs of the first n search result hits are collected and recorded. The number n may vary. The distillation algorithm may work even with a minimal number of search result hits such as for example 10. Subject to resource and time constraints, there is of course no limit to the number of search result hits that can be processed. However, processing more than 100 search result hits does not appear to significantly improve the confidence level of a matched or detected website address. Duplicate URLs in the set of search results are of course not counted twice.

For the URLs in the first n search result hits, a HTML spider is used to download the web content at each URL. The downloaded web content is parsed for website addresses and email addresses, which are compiled as follows:

For the 1st URL, for example, at www.somesite.com, the following email and website address are identified:

bob@company1.com
fred@company1.com
www.company1.com
sarah@company2.com
www.company2.com
www.company2.com
www.company3.com
bill@company1.com

The website addresses and email addresses are counted as follows for the 1st URL:

bob@company1.com is an email address and one count is added for website address "companyl.com".

fred@company1.com is an email address, however, since it has the website address "companyl.com", it is considered a "duplicate" website address and is not counted again,

www.company1.com is a website address and another count is added for website address "companyl.com".

In summary chart form, the email and website addresses associated with the 1 st URL are compiled as:

bob@company1.com	Email	+1company1.com
fred@company1.com	Email	Duplicate
www.company1.com	Website	+1company1.com
sarah@company2.com	Email	+1company2.com
www.company2.com	Website	+1company.com
www.company2.com	Website	Duplicate
www.company3.com	Website	+1company3.com
bill@company1.com	Email	Duplicate

For 2nd URL, the following email and website addresses are identified:

mrsmith@newfirm1.com

mrjones@newfirm2.com

www.company2.com

www.newfirm2.com

The email and website addresses associated with the 2nd URL are compiled as:

mrsmith@newfirm1.com	Email	+1newfirm1.com

mrjones@newfirm2.com	Email	+1newfirm2.com
www.company2.com	Website	+1company2.com
www.newfirm2.com	Website	+1newfirm2.com

For 3rd URL, the following email and website addresses are identified:

www.company2.com
www.anotherfirml.com

The email and website addresses associated with the 3rd URL are 20 compiled as:

www.company2.com	Website	+1company2.com
www.anotherfirm1.com	Website	+1anotherfirm1.com

For 4th URL, the following email and website addresses are identified:

mrbrown@newfirm3.com
mrjjones@newfirm2.com
www.company2.com
sarah@company2.com
www.anotherfirm2.com

The email and website addresses associated with the 4th URL are compiled as:

mrbrown@newfirm3.com	Email	+1newfirm3.com
mrjones@newfirm2.com	Email	+1newfirm2.com
www.company2.com	Website	+1company2.com
sarah@company2.com	Email	+1company2.com
www.anotherfirm2.com	Website	+1anotherfirm2.com

This processing continues for each URL of the first n search result hits.

Processing Matches

After each URL has been compiled as noted above, the compiled results are added to a master table to create running totals as follows (assuming 4 URLs have been processed):

	Emails and Websites	Websites only (n=4)
Company1	2	1
Company2	6	4
Company3	1	1
Newfirm1	1	0
Newfirm2	3	2
Anotherfirm1	1	1
Newfirm3	1	0
Anotherfirm2	1	1

After processing 20 URLs, the running totals may be

	Email and Websites	Websites only
Company1	3	1
Company2	28	19
Company3	1	1
Newfirm1	1	0
Newfirm2	8	3

Anotherfirm1	1	1
Newfirm3	3	0
Anotherfirm2	1	1
...		
Newfirm_x	2	1

In the (n=4) running total example, the highest value 6 for company2 is double that of Newfirm2. In the (n=20) running total example, Company2 has over three times the count of the combined total (Emails and Websites) and over six times the total of the Website only count as that of Newfirm2.

The prediction algorithm may be set to deem a match for a website address to be that of an entity when the highest count for a particular website is a multiple of the second highest count after processing a minimum number of x search result hits. As n increases, this ratio will also likely increase. Thus, processing of the search result hits may also stop after n (> x) URLs are processed when the prediction criteria for a website address determination is satisfied.

In a search for a business' website address, there may be cases where, for example, two. website addresses have similar counts as shown in the following example:

	Emails and Websites	Websites only
Company1	3	1
Company2	28	19
Company3	1	1
Newfirm1	1	0

Newfirm2	22	15
Anotherfirm1	1	1
Newfirm3	3	0
Anotherfirm2	1	1
...		
hotmail.com	24	0
Newfirm_x	2	1

In this case, Company2 and Newfirm2 are both considered to be matches for the website address of the business. There may be a number of reasons for this situation such as for example, the business uses 2 URLs for their website, one URL was previously used but has been replaced and another URL is now being used, or that one URL is a false match and is actually a directory or news site. Known directories or news sites may be designated as false positives and be removed.

Prediction Techniques

According to a further embodiment of the present invention, the prediction engine may be set to determine a match when a website address has a number count that is a multiple of either the mean or median count after processing a minimum number of x search result hits. Thus, both the Company2 and Newfirm2 website addresses may be identified as the website addresses of the business.

According to a further embodiment of the present invention, the distillation algorithm may further include the step of verifying a website address by matching further attributes of the business, such as for example the business name and address, to the content of the website linked to the website address. This feature is particularly significant when one or more of the search engines return only a few search result hits. This could be due to a number of reasons including there is no website for the business, the website is not well represented in search engines, or the website is not well linked to/by other websites.

In these cases, a clear pattern may not be established from the search result hits, such as for example, the search results may yield only three or four possible hits and/or a small number of URLs. In this case, the master table may include a list of website addresses, all with an associated count of 2 or 3. Rather than identifying all of the website addresses as possible matches, the websites linked to each website address in the list is searched for the physical address and business name of the business of interest. For example, assume "Bob's Pizza, 123 Main Street, Chicago" is submitted, a telephone number of 123-555-1212 is returned, and the following five potential matches are identified in the search results:

URL_A
URL_B
URL_C
URL_D
URL_E

Each of the potential matches, URL_A to URL_E is visited and searched for physical addresses. If only one physical address is found and it is 123 Main Street then this URL is deemed to be a positive match. If several physical addresses are found, but only one of the addresses is 123 Main Street, then this URL may be a match, but it could also be a directory. If one or more physical addresses are found, but not 123 Main Street then the URL(s) is not considered to be a match. Software to locate physical addresses including addresses in graphics that are on web pages is known.

In addition, if any of the physical addresses on the web pages matches an address that is known NOT to be Bob's Pizza or the URL is known to be a directory or portal, then the prediction engine may be set to reject the particular URL in question.

According to another aspect of the present invention, a method to create and update a database of directory websites that include directories, news sites, or portals is provided. These directory websites are websites that display multiple addresses of other businesses in the regular course of business such as a Yellow Page directory, or newspaper site reporting news, or a local city portal.

Referring now to Figure 2B, the method comprises the steps of sending a large number of known entities to a search engine to yield a set of search result hits (step 200), receiving the search results (step 210), and correlating the search result hits into a matrix. The URLs returned in the search results are identified on the X axis and the known entities are identified on the Y axis (step 220). The known entities, such as businesses, do not have to have a website for this method to work.

URL Matrix

Referring to Figure 3, there is shown a sample matrix generated by the method of Figure 2B. The matrix shows a number of URLs that appear to be linked to websites having the content of several different businesses or entities. Some of this may be innocent such as similar names or telephone numbers, but URLs to directories, such as Yellow Pages, portals or news sites tend to yield many more hits and thus, stand out as directories for easy identification by software. As shown in Figure 3, URL #3 and URL #11 may be easily identified as directories.

For example, a local restaurant portal that lists 100 restaurants in a city would match for one hundred different restaurants. If the known businesses along the Y axis contained even as little as ten of these restaurants, this website would stand out as a directory and would be automatically added to the database of directory websites. Likewise, a news site such as the Washington Post may frequently include articles on particular businesses, and thus would also stand out during this process. This method uses large multiple hits/matches above a certain threshold for a website to be classified as a directory website. The selected threshold depends on the sample size and may be any positive number above 2.

Directory of Websites

In a further embodiment, a database of directory websites is created in accordance with the present invention, using in particular the processes illustrated in Figures 2 and 3, to provide an index of directory websites that can be queried to locate directory websites or certain types of directory websites. It will be understood by those skilled in the art that the present invention may be modified to also rate directory websites by subject matter content. For example, a directory may be rated by the number of hits according to restaurants, types of restaurant, and locations of restaurants in its database. Thus, a user desiring access to a directory with restaurants in New York may be provided with a list ranked accordingly. The top restaurant directory website would be the one with the most hits of a sample set of restaurants from New York by that directory website.

Locating a URL

Figures 4A and 4B are flow diagrams of a method for locating a website address of an entity in accordance with a further embodiment of the present invention. The steps in this embodiment include selecting an input attribute, that is not a URL, but identifies an entity or business, by telephone number, physical address or business name (step 400); collecting other associated attributes associated with the input attribute (step 405), for example a telephone number may be associated with a physical address, which can be obtained from a standard telephone record database; submitting a query to one or more search engines, and any other databases of indexed content, using the input attribute and one or more of the associated attributes (step 410); receiving search results from the search engines and databases where each search result hit consists of a header, brief text description, and URL as well as possibly other information that may be provided (step 415); removing from the search results all URLs which are known to be associated with entities that are not the entity described by the input and associated attributes (e.g. directories, news sites, local portals) (step 420); and providing a no website address located answer if the number of search result hits is below a minimum number n (step 425). If n=0 then the entity is categorized as having no website.

Otherwise if n > 0 but below a minimum value , then the entity could be categorized as either having no URL associated with it; a low percentage of likelihood of the entity having no URL; or indeterminate. It will be understood by one skilled in the art that one of these actions may be chosen based on a number of different factors including personal preferences or past results as indicators of the likelihood of future occurrences.

If the search results yield a number of hits greater than or equal to n then the brief text description is analyzed (step 430), where provided (for example, Google), which includes the text immediately preceding and occurring after the matching text of the query attribute, as the actual indexed content ofthe web page. This obviates the need to download the content of such URL for further analyzes and recognizes that even if the number of hits is greater than n, if the brief text descriptions do not provide conclusive matches, the process can continue.

After step 430, the following steps are performed: downloading the web pages referenced by the first x number of URLs of the search result hits starting with the highest ranking URL (step 435); collecting from these web pages email addresses and website addresses (step 440); performing at least one of the following: limiting the matching of email addresses and website addresses to a set distance (in ASCII characters) prior to or after the matching attribute; limiting the number of matches of any one website address or email address to a count of two (once for a website and once for an email) so that one URL that lists the same website address or email address several hundred times does not negatively bias the results; and removing email addresses relating to public email services such as HOTMAIL (step 445); compiling the website addresses and email addresses from the downloaded web pages and accumulating a running total of all of the collected email addresses and website addresses (step 450); examining the totals of each website address and email address collected both individually and by combining emails addresses and website addresses that have the same primary and secondary domain (for example, www.geosign.com and timnye@geosign.com are the same) after each URL has been downloaded (step 455); determining one or more website addresses for the entity using a prediction algorithm where the algorithm matches a website address when any one total is greater than the next nearest total by a factor of N1 where NI can be any positive number greater than 1 or is greater than one of the average/median/mean number of matches per URL by a factor of N2 where N2 can be any positive number greater than I (step 460); and if no total is above N1 or N2 and if there are more search result hits to process then repeating steps 430 to 465 using the next URL in the set of matching search results (x) where x is set to 1 from the original query matching URL or using the next x number of URLs where x is set greater than 1. If at least one total is above N1 or N2 and n number of search result hits have been processed, then the matched website address(es) is provided. If all of the search result hits have been processed and no total exceeds NI or N2 then the original entity is categorized as either having no URL associated with it; a low percentage of likelihood of having no URL associated with it; or indeterminate.

It is anticipated that N1 and N2 will likely be in a range of greater than 400% or a factor of 4 from test results to date where there are a large number of search result hits. The idea is that with several hundred samples, one or two website addresses will stand out as spikes in an X/Y graph as shown in Figure 5, which illustrates a graph of counts (occurrences) versus URLs (websites and/or emails) of a sample result, where the main spike (n=18) is most likely to be the website address and the secondary spikes (n=5) and (n=6) are likely to be portals or directories.

When the number of search result hits is very small (total less than 20), then there may be website addresses with counts of 2 or even 1. Thus, further criteria for the prediction module may include a minimum number of search result hits for a match to be determined.

Implementation Examples

Applications making use of the present invention will now be described for illustrative purposes and are not intended to limit the scope of the invention.

Various devices may be used to input an attribute at step 400 in Fig. A, such as an application running on a portable device. For example, the device may be a RIM pager or Palm Pilot running a program such as Vindigo, that provides address information about businesses near a user, using some form of menus or categories. The user may desire to obtain information about a business within a certain distance from his location. However, the information provided to him on that business is usually just the location of the business on a map, an address, and possibly a telephone number and/or some other basic attributes. According to the present invention, the user may identify any point of data displayed on the device using a variety of programmable methods (mouse, stylus, voice, touch) and request more information on the identified point of data. The data identified may be linked to a telephone number (or submitted as is) and from this a website address is determined. Data is downloaded from the website and presented to the user.

Further, a smart agent may be used to analyze the downloaded data prior to displaying it to the user in order to anticipate the information that may be of interest to the user. For example, the smart agent may parse the contents of a restaurant website for menu descriptions and return a query to the user asking if the user would like to view the menu. Alternatively, the smart agent may analyze the menu to determine if the restaurant is a low priced restaurant or high priced, and thus, determine if the user would enjoy the restaurant or not.

For a clothing store, the smart agent may search for certain brands that 10 the user may have previously indicated an interest in, or find general specials to present to the user.

Further, the user may not even have to select the data point but rather may use a communication device, which is in the user's possession such as one built into a car, a cell phone or other portable device that has some GPS or positioning ability. In this case as the user moves around, the local entities in the area are located by a database of telephone numbers or other attributes, and the contents of their websites are downloaded on the fly and presented to the user, or processed at some location so that when the user performs a query, the local data is already freshly indexed. Thus, the user may always have all Internet content within a set range (eg. 10 miles) available either locally in their communication device, or on a central server, which can easily be queried by the user. As will be appreciated, this process saves a large amount of query time when the user needs local information. This also ensures that the information is current. Currently, queries to a search engine are only as fresh as the latest update or spider performed by that search engine, which may be good for some websites, poor for others, and non existent for others.

In a further embodiment, a user may provide an attribute, such as a telephone number, over a wireless telephone device so that content from the website such as menu information, or store specials, are provided via a WML browser (if their device and the website are so compatible) or by reading the text using common text to voice technology.

In a further embodiment, an intelligent web agent may read the web content linked to a URL in real time and intelligently construct an option to a user based on the read web content. For example, if a user was to ask for the telephone number of a restaurant, the distillation algorithm may determine the URL, read the web content and ask the user, "Would you like to hear/access their menu?". If the query was for a department store, or a clothing store, the question generated might be "This store has a sale today on ProductX. Would you like to order one?" Note that in this second case, the process is further enhanced as the intelligent agent is able to recognize the online ordering process for the business and cross reference that with the web content so that the user can actually interface with the website intentionally.

In a further embodiment, a rating system is provided based on the concept that whatever content is on a web page is irrelevant and not known ahead of time. Most systems that use an interface rely on known datasets. A smart agent may also generate time dated comments such as "This business has not updated its website in over six months", (by examining when web page was last cached). This fact could be used on its own or combined with other generated facts from both online and offline businesses to provide a rating for a store, so that stores with high ratings could be queried. This would improve customer service, lead to faster web updates and lower prices as user feedback would drive businesses to be more competitive.

It will be appreciated by those of skill in the art that the source of input language is irrelevant. Any attribute provided by the user can be linked to a telephone number and therefore, as numbers have no language dependence, can be linked to a website that may contain content in any language. This content may be read back to the user in the original language of the user or in the language that the content is written in, or in fact any language. The ability to read back the web page (deliver the content of the website) in the same language as the user is accomplished by determining the language of the user initially. This can be done very easily if the user says a telephone number using a language database capable of recognizing numbers in several languages.

Alternatively, this also could be accomplished through user input. The user may be asked to select a language (one for English, deux pour francais) and the selected language recorded. Once the query is made by the user (attribute is supplied), the query is matched to a telephone number using either automated or human methods, and from the telephone number the website is located using one of the methods described previously. Once the website is determined, using the intelligent agent, the web content is read back to the user using a text to voice program. In accordance with the present invention, an attribute may be received via voice or Internet and in response, a website returned by either looking the website up in a database associated with that attribute or by performing a real-time process such that the website address is determined from the attribute in accordance with one of the above described methods. For example, a query for a website address is looked up in the database: if not present or if present, but the data is determined to be stale (i.e. the website was last updated beyond a certain time period), then the website address is obtained in accordance with one of the above identified distillation techniques and the database is updated; and otherwise provide the website address in the database. Thus, the currency of such databases is maintained since they are updated.

In a further embodiment, a user is also asked if they would like the website of a business when they request the telephone number of the business (such as from 411 or telephone directory assistance), either as a free service or at a charge.

In a further embodiment, a user is asked questions based on the calling device used. The means to recognize a calling device if a 3G enabled telephone or a user calling using a computer headset on a PC, or whether the telephone has a color display or is a hybrid telephone/personal assistant type device are known. Further, the user is presented with different options based on his likely actions. For example, a user with a RIM pager would be offered "Press 7 to add this information to your address book. There will be a 75 cent charge for this service". A user with a 3G color telephone who is calling about the nearest theatre would be offered "Press 7 to view a trailer of the current movies showing now.". This feature would not be offered to someone calling on a normal telephone which cannot display video.

In a further embodiment, the content from a website, or other content, may be downloaded into the memory or hard storage in the user's calling device for offline viewing. This may be with or without the user's knowledge or consent and this content may be timer or location evoked to trigger a future action. For example, a user calls an 'information service' and asks for the number for "Pizza Hut" using a 3G telephone with the ability to run applets. The telephone number is provided and the user may be offered various choices. While online, the present invention recognizes the word 'Pizza' or even the firm 'Pizza Hut' and downloads an applet to the user's device containing a smart advertisement (or other actionable item). Two days later the user walks past a pizza store between 11:00am and 1:00pm that is having a special event. The applet triggers an event and beeps. If responded to, a message that pizza company X which is near by has a 2forl special on now is provided to the user. The user may or may not be aware that the applet was installed on their device. As will be appreciated, the point of contact with the user is the information service, that the applet downloaded was directly related to the query made by the user and that the applet can contain a number of triggers such as time and location.

In a further embodiment, a user uses a telephone to dial a telephone number (eg: 1-800-website) for automated access where the user could then type in the telephone number of the business or speak the telephone number into the telephone and have it converted, and then the user would be taken to that website and be able to have parts of the website read back either on mass, or using an intelligent agent or some menu system. Alternatively, the URL may be returned to the user for subsequent web access upon further actions from the user. In a further embodiment, portions of a website's HTML content is read out to the user including title tags or keyword tags as a way of confirming that this is the correct site. For example, a response from telephone: You asked for "THE EMPORIUM". In Boston, we have 2 matches. One is about food, pizza, delicatessen, coupons accepted, the other is about "we carry all XXX movies and toys". Press 1 for first or 2 for second.

In a further embodiment, an audio tag is defined on a website, which could be read to a user accessing the website via an audio device. For example, using the above "The Emporium", the message read back for the first Emporium match as follows: "Come visit The Emporium, Boston's finest delicatessen and home of the Dagwood. Located in Market Square inside the Marriort Hotel".

In a further embodiment, a website selects, for example based on user preferences, a voice from a library of choices by using HTML tags <tag audiotag voice="Female Serena" Content ="Come to Bigboy's steakhouse where we will treat you right!"> or other method.

In a further embodiment, the voice used to read text to a caller has the same accent or ethnic background as the caller but not identical so as to be mimicking. The accent or ethnicity may be determined by analyzing the caller's initial voice query so as to provide a more positive customer experience and to ensure clearer communications as people tend to understand better the speech of others with the same accent or ethnicity.

In a further embodiment, a user inputs via a telephone or a web interface and provides an attributes) describing a business (e.g. name of business and city). Currently the normal process provided by a 411 service is to offer the telephone number, and some 411 services connect the user directly by offering to dial the number for a fee. The present invention enhances this 411 system by offering the user the ability to obtain the website of this business (e.g. "Press 9 for the website of this business"). Currently, the only technical way to do this is to have a database of websites and telephone numbers or business names, and perform a lookup. Unfortunately, these databases are not available today in any complete form. In a further embodiment, the database of websites is created using the present invention as described above. In a further embodiment, this invention may be used in addition to a database to verify all answers from the database prior to submitting them to a user.

It will be understood by those skilled in the art that the use of email addresses is not necessarily required to practice the present invention, but is an improvement to collecting just website addresses. The collection of email addresses in addition to website address provides a greater confidence level to any determination of a website address.

In a further embodiment, the present invention may also be applied to a collection of email addresses. Thus, a website address of an entity is first determined, then all of the emails that were collected in the process of determining the entity that had the same website address are returned. Thus if a telephone number 123-456-7890 returned WWW.BUSINESSONE.COM as the website address, then BRIAN@BUSINESSONE.COM and FREDC@7BUSINESSONE.COM are considered to be email address matches.

It will be understood by those skilled in the art that the present invention may also be used to collect various other attributes associated with the website once the website is identified.

In a further embodiment, the prediction module may be based on a co-efficient (or threshold value) defined as the total matches of an individual URL divided by the number of matches to the original query, where correct matches exceed a certain coefficient value. The coefficient value may be determined by setting a value, which includes all or most of a set of known correct matches.

As will be appreciated the present invention provides advantages in that the content being provided is not taken from an out-of-date directory type database, but rather is fresh and up-to-date, taken directly from the business' website or from an updated directory.

Thus if a user using a program such as Vindigo on a Palm or other supported wireless device requests a list of all restaurants with a 4 star rating within 5 miles of them, the search results are displayed as a list of restaurants meeting the criteria. The user touches the name for `Restaurant A' and says `web' (or presses f5 for example). The program evokes one of the above described methods, which first checks to see if the search result hit is already in the database, and/or otherwise performs a realtime lookup to locate the website, and then if the user's device supports web browsing, loads www.restaurants.com or otherwise returns the URL linked to Restaurant A's website.

Alternatively, the process allows the user to pull up a box that allows them to search the website for a particular string if they do not have web browsing ability. The ability to do this already exists on the web (google plugin) but requires the user be on the Internet to do this. This invention is an improvement in that it allows the user to perform this without being online or even knowing the website.

In addition, the user can highlight several displayed entities, and ask for the list to be filtered by a particular keyword. For example, the user highlights 10 seafood restaurants and wants to see which ones serve "sea bass". The system locates the websites, searches them for the words "sea bass" and then returns the matches in some form of user interface.

Regardless of the whether the user actually selects a telephone number or an entity, or is simply looking at a map and points at an icon on the map, the present invention attaches attributes to that icon, which may be an entity name or telephone number, or that the entity name may in turn have an attribute of a telephone number. This enables the process of going from icon to entity to telephone to the distillation module to web content (or to any attribute or information requiring web) or the process of going from icon to the distillation module directly and to web content. The whole process can be assigned to a single key (but not limited to) or voice command.

A string of text or voice can also be parsed for semantic meaning and/or a one word input can be used to look up all the matching entities (assuming that the geographical location is known) in the current online Yellow Page listings. The group of telephone numbers can then be used to create a group of websites and a response back can be formulated based on querying of these websites.

For example, a user says "tell me about the restaurants in town" or says "restaurants" and from the wireless device location we know the user to be located in downtown Toronto at a particular latitude and longitude. The system looks up all the matches it has for restaurants and returns a set of names and telephone numbers. If websites are known for all these businesses then the distillation process need not be used (covers the case where all websites are known), otherwise the distillation process determines the websites for these businesses, and could even add to this list if other matches are found (covers merging multiple databases of websites). When a set of websites is located (not all may have websites) the content of the websites is downloaded into memory and processed with some form of avatar process to provide an intelligent user response based on the content contained on the websites, which could be based on a number of factors. This experience can augment any system. The user is then able to interact with the website content of the restaurants through user prompted questions or free flow questions depending on the level of available semantic processing.

Example of the process described above using restaurants is set out below: (Text in italics represents what cannot be done without the web, normal text represents what is easily done now, bold text represents user input shown as single words for brevity but these could be mouse clicks from a menu shown to the user, or phrases interpreted by a semantic parser). Specifically, two examples of queries are described below with the queries shown in quotations and the responses provided below the queries.

OPTION 1

"Restaurants"

There are 17 restaurants within 5 miles of you. *Six restaurants have specials fo*r *tonight, 3 restaurants have live entertainment;* four restaurants require reservations.

"Italian"There are 3 Italian restaurants. The closet one is 500 yards. One requires a reservation and also has *live entertainment.* The names are Olive Garden, Pasta Mania, La Tratoria. *The most popular restaurant appears to be Olive Garden.*

"Olive Garden"

The Olive garden is 1.5 miles away, *and has a `Wines of Northern Italy" promotion going on right now. Kids under 10 eat free.* Say 'menu' to hear/see the menu. The price is medium.

"Menu"
They have six categories, Entrees, Appetizers, lighter fare, House Specials, Desserts and drinks.

"Entrees"
The special today is "Angel hair Primavera"

And so on.

OPTION 2

"I want a steak house with a nice selection of wines?"

There are four steakhouses within 10 miles of you. *Mortons lists wines from Chile,* California, *and France,* Ruth Chris is also a four star restaurant, *but does not list its wines. The other two restaurants are not licensed.*

"Tell me more about Ruth Chris's"

The Ruth Chris restaurant is located 1.6 miles by car from your present location, and is a four star restaurant located in the lower level of the Hilton Hotel. *There are 36 locations across the US. The restaurant's claim to fame is that it features the finest steaks broiled to 1800 degress cooked in butter to maintain the corn fed* flavor. The restaurant has had favorable reviews from

Fodors, and AAA. Reservations are recommended. *Would you like to hear the menu?*

"Yes"

There are five choices of steak, T-bone, New York, Ribeye, Filet, Porterhouse. All side dishes are extra. They do not guarantee steaks that are well done.

"Book"

I am attempting to book this using the online reservation system. How many people and what time would your prefer please.

"Four people for 7:00".

Thank you. I will page you when I have a confirmation for you.

The above described process is applicable to any web content and not just restaurants. Programs can be written to intelligently parse content by vertical market, so that when a user requests information on restaurants, a special program or avatar is evoked to process this information and a different program is used if the consumer requests shopping information.

Accessing Up-to-date Content

The ability to use web content to provide a better user experience is unique. Systems that provide services like this currently use existing databases, and do not rely on using the web. These services are never complete as at any given time, a business may change its information and this change is only likely to be reflected on its website. Systems that use databases are also biased on the content that they have. The present invention, however, uses web content, and therefore is only biased by who does not have a website. This situation is becoming rarer. Systems that use databases face huge daunting tasks of maintaining their data current.

Database Integration

There are a wealth of databases containing information that can used in a search routine. These databases may be mailing lists, memberships lists, etc. that all share a commonality in that they are all collections of data. Examples of collections of data are members of the Better Business Bureau, members of the AARP, merchants that take Visa, doctors, or gas stations that take diesel. Such a collection of data may be used to create an enhanced search experience for the user.

One example is using a list of doctors to determine whether any of the listed doctors makes house calls. The list in this example contains the names, addresses and phone numbers for all the doctors in the state. The user, via an interface, indicates that he or she wants to find a doctor that makes house calls. This could be input at a search engine interface. The system may use the phone number of each doctor to determine the URL of the doctors in that state. Then, the system may go to each URL and look for the phrase 'house calls' or 'we do house calls' and return the results that match the user's query. By initially providing a list of doctors, the system can ensure that any matches are at least doctors from the list. By way of contrast, a search on a generic search engine might return listings for a TV station selling a comedy called 'house calls' or a medical journal discussing the effectiveness of 'house calls'.

Another example is a user that has a database of hotels rated 3 stars or above by AAA. This database of hotels may be spidered (collect the data) and indexed. The database may or may not include the URLs for the hotels. The resulting index would be useful for a travel search engine to filter its search results through. For instance, executive travelers, could make queries such as 'pool', 'day care', 'high speed internet access' knowing that all the results are hotels, and there are no mismatches from outside this list of hotels. The system could identify the URL addresses associated with each hotel, and determine whether any of the hotel's websites match the user's search query.

Determining URLS based on information from a database can be performed by cross-referencing the database against another collection of data, including information about businesses, such as phone numbers. In this way, a database of records can be used to determine URLs, even though the database may not necessarily have URLS as attributes. The content of the websites at the URLS may be indexed and used to determine whether any of the URLs match the user's criteria. Thus, highly targeted searches may be achieved by cross-referencing and narrowing search results using this collection of information.

Search Filters

According to an embodiment of the invention, a library of search filters (or parsers) may be used to focus search results in real-time. Each search filter may correspond to specific subject matter. For example, a restaurant search filter may include a specialized parser for restaurant related data. The user may type in 'Italian food' as the query and instead of looking for the words 'Italian food', a parser might look for words such as 'pasta, linguine, lasagna' and return matches for all URLs that contain these words.

According to an embodiment of the invention, a database is selected for the user based on the user's query. For example, if a user inputs an "Italian Restaurants" query, a database may be selected that reflects the query. In this example, an appropriate database may be a restaurant database. A restaurant database may be generated, for instance, by extracting a list of restaurants from a Yellow pages directory of restaurants. The URLs for the restaurants may be determined, and then a search for Italian food may be performed on the website associated with each URL. A similar technique which uses the contents of a database as a geographic location filter to a query interface is described in U.S. Application No. 10/620,170, the entire teachings of which are incorporated herein by reference.

Figure 6 shows the process for using a database as a filter for a search routine according to an embodiment of the invention. At 600, a user inputs an attribute of as a search query. The attribute may be a phone number of a business, a phrase (e.g. "Italian food"), etc. At 605, the search query is received. At 610, the system determines whether a database has been identified. A particular database may, for example, be selected by the user. If a database has not been selected, then at 615 the system chooses an appropriate set of records that reflect the query received from user. At 625, the determines candidate URLs. The URLs can be determined by database lookup or by using an algorithm. At 630, the user has the option of receiving the potential URLs so they can visit the website on their own. Otherwise at 640, the system collects the data from the websites associated with the potential URLs. This can be performed by spidering the website and collecting raw data. At 645, the system may collect data from other webpages associated with the URL's domain name. At 655, the system parses the website data, based on the user's query and returns the results. At 660, the results are compared with results from a search engine using the query. The search engine could be an existing search engine, such as Google or Yahoo. Any matches identified in both search results are returned to the user. Computational techniques may be performed on the results to eliminate false positives and to determine the most likely match.

While this invention has been particularly shown and described with references to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.